تشخیص داده‌های دورافتاده دایره‌ای با استفاده از یک مدل آمیزه‌ای از توزیع‌های فون‌میزس

خدیجه عبدی، موسی گل‌علی‌زاده[۱] و تابان باغفلکی

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۵/۱۱/۶ تاریخ پذیرش: ۱۳۹۷/۲/۵

چکیده: داده‌های دایره‌ای نوعی از داده‌های جهتی با دوره تناوبی مشخص هستند. به دلیل اینکه وجود داده‌های دور افتاده استنباط‌های آماری راجع به پارامترهای مدل‌های رگرسیون دایره‌ای را نامعتبر خواهد کرد، بررسی وجود آنها در تحلیل این مدل‌ها نیازمند توجه ویژه‌ای است. روش‌های متنوعی برای مدل‌بندی ساختار مجموعه داده‌ها شامل مشاهدات دور افتاده وجود دارد که به‌کارگیری مدل آمیزه‌ای یکی از مهم‌ترین آن‌هاست. در این مقاله علاوه بر مطالعه یکی از روش‌های مدل‌های رگرسیون دایره‌ای، به‌عنوان ایده پژوهشی جدید نحوه به‌کارگیری مدل آمیزه‌ای از توزیع‌های فون‌میزس برای تعیین داده‌های دورافتاده در این مدل تشریح می‌شود. برای برآورد پارامترهای مدل مذکور از الگوریتم EM استفاده خواهد شد. عملکرد مدل پیشنهادی با استفاده از مطالعات شبیه سازی بررسی و سپس از ماتریس درهم ریختگی برای ارزیابی اعتبار برازش مدل کمک گرفته می‌شود. به‌علاوه، روش‌های پیشنهادی در تحلیل داده‌های جهت موج سواحل استان بوشهر مورد استفاده قرار می‌گیرد.

واژه‌های کلیدی: داده‌های جهتی، نقاط دور افتاده، توزیع فون‌میزس، مدل‌های آمیزه‌ای، الگوریتم EM.

رده‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۲J۲۰.

۱- مقدمه

یکی از موضوعات مهم علمی موردتوجه محققین علومی مانند زیست شناسی، زمین شناسی، هواشناسی و فیزیک، مطالعه بر روی داده‌های جهتی است. جهت‌های دو بعدی که به‌صورت زوایا نشان داده می‌شوند، با توجه به جهت صفر یعنی نقطه شروع و جهت دوران یعنی موافق یا

۱- آدرس الکترونیکی نویسنده مسئول مقاله: golalizadeh@modares.ac.ir

مخالف بودن آنها با حرکت عقربه‌های ساعت اندازه‌گیری می شوند. ازآنجاکه جهت‌ها فاقد مقدار هستند، می‌توان آنها را به‌صورت نقاطی بر روی محیط دایره‌ای با شعاع واحد به مرکز مبدأ نمایش داد. به همین دلیل است که مشاهدات در فضای دو بعدی با طول واحد را داده‌های دایره‌ای[۱] می‌نامند. به دلیل ماهیت تناوبی بودن داده‌های دایره‌ای، باید تحلیل این نوع داده‌ها با روش‌هایی متفاوت از روش‌های معمول برای مجموعه داده‌های غیر دایره‌ای انجام شود. چنین تحلیلی در شاخه جدیدی از آمار با عنوان آمار دایره‌ای امکان‌پذیر ا ست. برای تفکیک داده‌هایی که از ویژگی تناوبی بودن در فضاهای نااقلیدسی برخوردارند، بعد از فعالیت [۱] مرسوم شده است که اصطلاحات خاصی متداول شود. به‌طور دقیق‌تر، آمار مرتبط با تحلیل داده‌ها در ف ضاهای برداری را آمار خطی و برر سی م شاهداتی در ف ضاهای نااقلید سی مانند دایره و کره را آمار غیرخطی می‌گویند [۲].

بحث راجع به مدل‌های رگرسیونی با متغیر پاسخ دایره‌ای اولین بار توسط [۳] برای تحلیل داده‌های دایره‌ای در حوزه علوم پز شکی مطرح شد. او توزیع فون‌میزس که معادل توزیع نرمال در فضای اقلیدسی است را برای تحلیل خطاها در فضای نااقلیدسی دایره معرفی کرد. [۱] مدل (۳) را به ازای مجموعه‌ای از مشاهده‌های وابسته دایره‌ای تعمیم داد. در ادامه این فعالیت‌ها [۴]، [۵]، [۶] و [۷] نیز مدل‌های رگرسیونی متفاوتی ارائه کردند. سپس، مطالعه جامع‌تری از ارتباط بین دو متغیر دایره‌ای توسط [۸] انجام شد. [۹] و [۱۰] این مدل را تعمیم داده و برآورد پارامترها را به روش ماکسیمم درستنمایی محاسبه کردند.

اگرچه مدل‌های رگرسیون دایره‌ای ارائه شده توانایی برازش به برخی داده‌های جهتی را دارند، اما وجود داده‌های دور افتاده از کارایی این مدل‌ها خواهد کا ست. ت شخیص داده‌های دور افتاده در مدل‌های رگرسیون دایره‌ای سابقه‌ای کم در آمار غیرخطی دارد. [۱۱] از نخستین کسانی بود که به برر سی چنین داده‌هایی در آمار دایره‌ای پرداخت و با معرفی چند آزمون آماری، راهکاری برای تشخیص داده‌های دور افتاده ارائه کرد.

[۱۲] با پیروی از رویکردی بیزی به تشخیص این نوع داده‌ها در آمار جهتی پرداختند. [۲] نیز رهیافت نوینی برای رویارویی با چنین داده‌هایی ارائه کردند. در ادامه چنین فعالیت‌هایی، [۱۳] نیز آزمونی دیگر پیشنهاد کردند و [۱۴] به مطالعه و تشخیص داده دورافتاده در مدل رگرسیونی [۹] بر اساس آماره نسبت دترمینان کوواریانس‌ها پرداختند. تعدیل آماره‌ای در آمار خطی، به‌منظور تشخیص داده دور افتاده در مدل رگرسیون دایره‌ای [۵] توسط [۱۵] پیشنهاد شد. در رویکردی دیگر، [۱۶] آماره میانگین خطای دایره‌ای را برای تشخیص این نوع داده در مدل رگرسیون دایره‌ای ساده معرفی کردند. در راستای این تحقیقات مقاله حاضر مدل آمیزه‌ای را در

۱- Circular data

نظر می‌گیرد که توزیع فون‌میزس نقش اساسی را در آن بازی می‌کند. به‌طور دقیق‌تر، فرض می‌شود که داده‌های دایره‌ای از مدل آمیزه‌ای از توزیع‌های فون میزس با درصد آمیختگی نامعلوم آمده‌اند طوری که پارامترهای توزیع نیز مجهول هستند. آنگاه، درصد آمیختگی و پارامترهای هر یک از توزیع‌ها با ا ستفاده از الگوریتم EM به صورت توأم برآورد می شوند. برای ارائه نتایج تحقیق، مقاله پیش رو به چهار بخش تقسیم شده است. بعد از مقدمه، در بخش دوم مدل رگرسیون دایره‌ای بر اساس بسط چند جمله‌ای مثلثاتی می‌آید. روش پیشنهادی این مقاله برای تشخیص داده دور افتاده در بخش سوم تشریح می‌شود. مطالعه شبیه‌سازی همراه با تحلیل یک مثال واقعی برای ارزیابی مطالب نظری ارائه شده در مقاله در بخش چهارم خواهد آمد. مقاله با نتیجه‌گیری کلی خاتمه می‌یابد.

۲- مدل رگرسیون دایره‌ای بر اساس بسط چند جمله‌ای مثلثاتی

کلیت [۵] برای دو متغیر تصادفی دایره‌ای x و y یک مدل رگرسیونی به‌صورت امید ریاضی e^{iy} به‌شرط x در نظر گرفتند، که در آن x متغیر تبیینی و y متغیر پاسخ بوده و $0 \le x, y < 2\pi$. این مدل، که آن را در این مقاله به‌اختصار مدل رگرسیون دایره‌ای JS می‌نامیم، به‌صورت

$$E(e^{iy} \mid x) = \rho(x) e^{i\mu(x)} = g_1(x) + i g_2(x) \qquad (1)$$

نوشته می شود که $\mu(x)$ میانگین جهتی y به شرط x و $0 \le \rho(x) < 1$ پارامتر تمرکز شرطی نسبت به این جهت است. به‌علاوه،

$$e^{iy} = \cos y + i \sin y . \qquad (2)$$

با اخذ امید ریا ضی شرطی (به شرط x) از طرفین رابطه (۲) و مدنظر قرار دادن عبارت (۱)، به دست می‌آوریم:

$$E(\cos y \mid x) = g_1(x), \qquad E(\sin y \mid x) = g_2(x), \qquad (3)$$

که با توجه به روابط مثلثاتی می‌توان نوشت:

$$\mu(x) = \begin{cases} \tan^{-1} \dfrac{g_2(x)}{g_1(x)} & g_1(x) > 0, \\ \tan^{-1} \dfrac{g_2(x)}{g_1(x)} + \pi & g_1(x) \le 0 \end{cases}$$

با توجه به‌دشـواری برآورد ناپارامتری $g_1(x)$ و $g_2(x)$ در (۳) معمولاً از تقریب چند جمله ای مثلثاتی استفاده می‌شود. مناسبت این تقریب برای برآورد، یادآوری این حقیقت است که هر دو تابع $g_1(x)$ و $g_2(x)$ متناوب با دوره تناوب 2π هسـتند. با پیروی از [۵] و اسـتفاده از چند جمله‌ای مثلثاتی می‌توان نوشت:

$$g_1(x) \approx \sum_{k=0}^{m} (A_k \cos kx + B_k \sin kx)$$
$$g_2(x) \approx \sum_{k=0}^{m} (C_k \cos kx + D_k \sin kx)$$

در نتیجه، با توجه به تساوی‌های ارائه شده در (۳) می‌توان فرض کرد:

$$\begin{aligned} \cos y &= \sum_{k=0}^{m} (A_k \cos kx + B_k \sin kx) + \varepsilon_1 \\ \sin y &= \sum_{k=0}^{m} (C_k \cos kx + D_k \sin kx) + \varepsilon_2, \end{aligned} \qquad (۴)$$

که در آن $\varepsilon = (\varepsilon_1, \varepsilon_2)^T$ بردار خطاهای تصادفی است که دارای توزیع نرمال با بردار میانگین $(0, 0)^T$ و ماتریس پراکندگی نامعلوم Σ است. یک مسئله مهم در برازش مدل رگرسیون چند جمله‌ای مثلثاتی تعیین درجه m است. برای این منظور، آزمون‌ها و معیارهای مربوطه باید طوری طراحی شوند که قابلیت بهنگام سازی برآوردها را با افزایش درجه چند جمله‌ای داشته باشند. در رگرسیون چند جمله‌ای معمولی برای متغیرهای خطی از چند جمله‌ایهای متعامد برای این امر استفاده می‌شود. وقتی متغیرها دایره‌ای هستند، روشی که بهترین درجه m را برای بهنگام سازی برآوردها بدون محاسبه تمام ضرایب جدید به دست آورد، بررسی کاهش مجموع مربعات خطا به ازای افزایش m است. برای جزئیات بیشتر به [۵] مراجعه شود.

به دلیل اینکه وجود داده های دور افتاده منجر به اسـتنباط های آماری نامعتبری راجع به پارامترها می‌شود، مطالعه آن‌ها موردتوجه محققین بی‌شماری بوده است. وجود چنین داده‌هایی در تحلیل آمار دایره‌ای نیازمند توجه بیشتری است چرا که مشاهدات دایره‌ای دارای ویژگی‌های خاصی مانند تناوبی بودن هستند. این در حالی است که داده‌های موجود در فضاهای برداری با این مشکل روبرو نیستند. روش‌های متنوعی برای مدل‌بندی ساختار مجموعه داده‌های دایره‌ای که شامل مشاهدات دورافتاده هستند وجود دارد. استفاده از توزیع‌های فون‌میزس آمیزه‌ای در داده‌های دایره‌ای به‌عنوان یک رویکرد جدید در این مقاله تشریح می‌شود.

۳-تشخیص داده دور افتاده بر اساس مدل آمیزه‌ای از توزیع‌های فون‌میزس

فرض کنید داده‌های در دســـترس در خوشـــه‌هایی مجزا قرار بگیرند. آنگاه داده‌های موجود در خوشـــه‌ای با حجم بســـیار پایین را می‌توان به‌عنوان داده‌های دور افتاده در نظر گرفت. در اینجا فرض می‌شـود که متغیر تصـادفی دایره‌ای θ از آمیزه دو توزیع فون میزس تولید شـده باشـد. درنتیجه، تابع چگالی احتمال آن با فرض اینکه $(\alpha_1,\alpha_2,p)^T$ که در آن $\alpha_h=(\mu_h,\kappa_h)^T$، $h=1,2$، بردار پارامترهای مدل باشد، به‌صورت

$$pf_1(\theta;\alpha_1^{(c)})+(1-p)f_2(\theta;\alpha_2^{(c)}) \qquad 0<p<1,0\le\theta<2\pi \qquad (۵)$$

خواهد بود که در آن $f_h(\theta;\alpha_h^{(c)})$ چگالی فون‌میزس با پارامترهای $0\le\mu_h<2\pi$ و $\kappa_h\ge 0$ یعنی

$$f_h(\theta_i;\alpha_h^{(c)})=\frac{1}{2\pi I_0(\kappa_h)}e^{\kappa_h\cos(\theta_i-\mu_h)}$$

است. در اینجا μ_h پارامتر میانگین، $\kappa_h\ge 0$ پارامتر تمرکز و $I_0(\kappa)$ تابع بسل تعدیل شده نوع اول، از مرتبه صفر است. برای جزئیات بیشتر [۱۷] را ببینید. برای برآورد پارامترهای این مدل آمیزه‌ای، از الگوریتم EM استفاده می شود [۱۸]. انتظار می‌رود این الگوریتم تفکیک داده‌ها در خوشـــه‌های مجزا و تشـــخیص خوشـــه‌ای با داده‌های دورافتاده را به‌طور مناســـبی انجام دهد. با پیروی از این رویکرد، معمولاً داده های داخل خوشه‌ای که شامل حداکثر ۵ درصد از کل داده‌ها باشـد به‌عنوان داده‌های دورافتاده در نظر گرفته می‌شـوند. در ادامه جزئیاتی از نحوه به‌کارگیری الگوریتم EM برای مدل (۵) به قصد برآود پارامترهای آن می‌آید.

فرض کنید $\theta_1,\ldots,\theta_n$ دارای توزیع به فرم (۵) باشـــند و به ازای هر یک از θ_i متغیر پنهانی به‌صورت z_i موجود باشد طوری که

$$z_i=\begin{cases}1 & \theta_i\sim f_1(\theta_i;\alpha_h^{(c)})\\ 0 & \theta_i\sim f_2(\theta_i;\alpha_h^{(c)}).\end{cases}$$

بنا به ادبیات الگوریتم EM به مجموعه داده‌های مشاهده شده و متغیر پنهان (اصطلاحاً گمشده) داده‌های کامل گفته می‌شـــود. نمادگذاری $\underline{\theta}=(\theta_1,\ldots,\theta_n)^T$ و $\underline{z}=(z_1,\ldots,z_n)^T$ را در نظر بگیرید. آنگاه، لگاریتم تابع درستنمایی داده کامل به‌صورت

$$\begin{aligned}In(\underline{\theta},\underline{z}\mid\alpha,p)=&\sum_{i=1}^{n}z_i[\log p+f_1(\theta_i,\alpha_1)]\\&+(1-z_i)[\log(1-p)+\log f_2(\theta_i,\alpha_2)]\end{aligned} \qquad (۶)$$

اســـت. با پیروی از [۱۸]، الگوریتم EM شـــامل دو قدم معروف گام E یا امیدگیری و گام M با بیشینه‌سازی است. برای تشریح جزئیات این الگوریتم، در ادامه مراحل محاسبه این دو گام برای تشخیص داده‌های دور افتاده دایره‌ای تشریح می‌شوند.

گام E: محاســـبه امید ریاضـــی رابطه (۶) یعنی $Q(\alpha,\alpha^{(c)}) = E\left(\ln(\underline{\theta},\underline{z}\,|\,\alpha,p)\right)$ توســـط تساوی

$$\begin{aligned} Q(\alpha,\alpha^{(c)}) &= \sum_{i=1}^{n} t_{i1}^{(c)}((\log p + \log f_1(\theta_i,\alpha_1)) \\ &+ t_{i2}^{(c)}(\log p(1-p) + \log f_2(\theta_i,\alpha_2)), \end{aligned} \qquad (۷)$$

صـــورت می‌گیرد که در این عبارت $t_{ih}^{(c)}$ احتمال شـــرطی $z_i = 1$ به‌شـــرط داشـــتن مقدار اولیه پارامترهای $\alpha^{(c)}$ و مشاهدات θ است. با توجه به محاسبات احتمالات پسین می‌توان نوشت:

$$t_{ih}^{(c)} = p\left(z_{ih} = 1\,|\,\theta,\alpha^{(c)}\right) = \frac{p_h^{(c)} f_h\left(\theta_i;\alpha_h^{(c)}\right)}{\sum_{h=1}^{k} p_h^{(c)} f_h\left(\theta_i;\alpha_h^{(c)}\right)}.$$

گام M: در این گام ماکسیمم‌سازی رابطه (۷) برحسب پارامترها مدنظر است. به زبانی دقیق‌تر، هدف محاسبه عبارت $\alpha^{(c+1)} = \arg\max\left(Q(\alpha,\alpha^{(c)})\right)$ است. در زیر نشان داده می شود که نتیجه این امر برای هر یک از پارامترها دارای فرم بسته است.

با مشتق‌گیری از رابطه (۷) نسبت به پارامتر μ_h به ازای $h = 1,2$ داریم:

$$\begin{aligned} \frac{\partial Q\left(\alpha,\alpha^{(c)}\right)}{\partial \mu_h} &= \sum_{i=1}^{n} t_{ih}^{(c)} \left\{ \frac{-\kappa_h(-\sin(\theta_i - \mu_h)) e^{\kappa_h \cos(\theta_i - \mu_h)}}{e^{\kappa_h \cos(\theta_i - \mu_1)}} \right\} \\ &= \sum_{i=1}^{n} t_{ih}^{(c)} \left\{ \kappa_h \sin(\theta_i - \mu_h) \right\}. \end{aligned}$$

با مساوی صفر قرار دادن رابطه اخیر، یعنی

$$\sum_{i=1}^{n} t_{ih}^{(c)} \left\{ \kappa_h \sin(\theta_i \cos\mu_h - \cos\theta_i \sin\mu_h) \right\} = 0$$

پارامتر μ به‌صورت

$$\hat{\mu}_h^{(c+1)} = \tan^{-1}\left(\frac{\sum_{i=1}^{n} t_{ih}^{(c)} \sin\theta_i}{\sum_{i=1}^{n} t_{ih}^{(c)} \cos\theta_i}\right) = \tan^{-1}\left(\frac{S}{C}\right)$$

به‌روز خواهد شد. با توجه به اینکه $\hat{\mu}_h^{(c+1)}$ میانگین جهتی ا ست، با پیروی از [۱۷]، باید برآورد آن به‌صورت

$$\hat{\mu}_h^{(c+1)} = \begin{cases} \tan^{-1}(\frac{S}{C}) & S > 0, C > 0, \\ \tan^{-1}(\frac{S}{C}) + \pi & C < 0 \\ \tan^{-1}(\frac{S}{C}) + 2\pi & S < 0, C > 0, \end{cases}$$

بازنویسی شود، که در آن $S = \sum_{i=1}^{n} t_{ih}^{(c)} \sin\theta_i$ و $C = \sum_{i=1}^{n} t_{ih}^{(c)} \cos\theta_i$.

برای به‌روزر سانی پارامتر κ_h نیز باید از رابطه (۷) نسبت به این پارامتر مشتق گرفت. می‌توان ملاحظه کرد که

$$\frac{\partial Q(\alpha, \alpha^{(C)})}{\partial \kappa_h} = \sum_{i=1}^{n} t_{ih}^{(c)} \left\{ \frac{e^{\kappa_h \cos(\theta_i - \mu_h)} \left(\frac{-2\pi I_1(\kappa_h)}{(2\pi I_0(\kappa_h))^2} + \frac{\cos(\theta_i - \mu_h)}{2\pi I_0(\kappa_h)} \right)}{\frac{e^{\kappa_h \cos(\theta_i - \mu_h)}}{2\pi I_0(\kappa_h)}} \right\} \qquad (۸)$$

$$= \sum_{i=1}^{n} t_{ih}^{(c)} (\cos(\theta_i - \mu_h) - A(\kappa_h)).$$

که به‌طورکلی $A_p(x) = \frac{I_p(x)}{I_0(x)}$ که در اینجا اندیس یک برای تابع A حذف می‌شـــود. با مساوی صفر قرار دادن عبارت (۸)، تساوی

$$A(\hat{\kappa}_h) \sum_{i=1}^{n} t_{ih}^{(c)} = \sum_{i=1}^{n} t_{ih}^{(c)} \cos(\theta_i - \hat{\mu}_h)$$

حاصل می‌شود که از طریق آن پارامتر κ_h به‌صورت

$$\hat{\kappa}_h^{(c+1)} = A^{-1}\left(\frac{\sum_{i=1}^{n} t_{ih}^{(c)} \cos(\theta_i - \hat{\mu}_h)}{\sum_{i=1}^{n} t_{ih}^{(c)}}\right),$$

به‌روز خواهد شد. پارامتر آمیختگی نیز مشابه دو پارامتر تمرکز و میانگین جهتی با مشتق‌گیری از رابطه (۷) به دست می‌آید. نتیجه حاصل به‌صورت زیر است.

$$\hat{p}_h^{(c+1)} = \frac{\sum_{i=1}^{n} t_{ih}^{(c)}}{\sum_{i-1}^{n} t_{i1}^{(c)} + t_{i2}^{(c)}} = \frac{1}{n}\sum_{i=1}^{n} t_{ih}^{(c)}\ .$$

اکنون می‌توان تشخیص داده دور افتاده با استفاده از روش الگوریتم EM را به‌صورت زیر تدوین کرد. به ازای تعداد نمونه تولید شده یعنی $i = 1,\ldots,n$ و دو خوشه $h = 1,2$ احتمالات پسین $t_{ih}^{(c)}$ محاسبه شود. سپس لازم است ماکسیمم مقادیر این احتمالات در هر مرحله محاسبه و مقادیر حاصل در یک خوشه قرار داده شوند. با انجام این کار، درنهایت دو خوشه به دست می‌آید. خوشه‌ای که داده کمتری دارد به‌عنوان خوشه دور افتاده شناخته می‌شود و داده‌های آن خوشه، داده‌های دور افتاده خواهند بود.

۴-مطالعه شبیه‌سازی و تحلیل مثال واقعی

در این بخش با انجام مطالعات شبیه سازی به بررسی روش تشخیص داده دور افتاده مبتنی بر مدل آمیزه‌ای از توزیع فون‌میزس با روش الگوریتم EM پرداخته می‌شود. صحت تشخیص داده دورافتاده با استفاده از این روش پیشنهادی با ماتریس درهم ریختگی بررسی می‌شود. علاوه بر این، تحلیل یک مثال واقعی مدنظر است. داده‌های این مثال مربوط به جهت موج و باد ثبت شده توسط بویه بوشهر است که از سازمان هواشناسی کشور تهیه شد. مدل رگرسیون دایره‌ای بر این داده‌ها برازش داده می‌شود و با استفاده از مانده‌های رگرسیونی، داده دور افتاده تشخیص داده می‌شود. تحلیل عملکرد روش نیز مورد بحث قرار می‌گیرد.

۴-۱-مطالعه شبیه‌سازی

برای انجام مطالعه شبیه‌سازی فرض می‌کنیم مجموعه داده‌ها از آمیزه دو توزیع فون‌میزس تولید شده‌اند. به طور دقیق‌تر، در نظر می‌گیریم داده‌ها از طریق رابطه

$(\pi,۱۲/۸)VM+۰/۹۵(\frac{\pi}{۲},۱۰/۳)VM+۰/۰۵$ به‌دست‌آمده باشند. در این حالت، می‌توان چنین فرض کرد که ۵ در صد از داده‌های تولیدی در یک خوشه و ۹۵ در صد از آنها در خوشه دیگر قرار دارند. لذا، مشاهداتی که در خوشه ۵ درصد هستند را به‌عنوان داده‌های دورافتاده در نظر گرفته‌ایم. برای تشخیص داده دورافتاده بر اساس مدل آمیزه‌ای از توزیع‌های فون‌میزس از روش الگوریتم EM که در بخش قبل تشریح شد، استفاده می‌کنیم. برای شروع شبیه‌سازی این مدل آمیزه، مقادیر اولیه زیر در نظر گرفته شد:

$$\mu_۱=۰/۵\pi,\ \kappa_۱=۱۴/۵,\ \mu_۲=۱/۰۱\pi,\ \kappa_۲=۱۱۶,\ p_۱=۰/۴۵$$

با این مقادیر اولیه، برآورد پارامترها به روش الگوریتم EM به‌طور متوسط در ۳۰ تکرار همگرا شدند. برآورد پارامترها به ازای حجم نمونه متفاوت و ۵۰۰ بار تکرار شبیه‌سازی در جدول ۱ آمده است. توجه شود که مقادیر داخل پرانتز، میانگین مربعات خطا (MSE) را نشان می‌دهد. همان‌طور که ملاحظه می‌شود با افزایش حجم نمونه، برآورد پارامترها با MSE کمتری به مقدار واقعی نزدیک می شوند. واضح است که چون خوشه با حجم ۰/۰۵ نمونه، تعداد کمتری دارد، برآورد پارامتر $\kappa_۱$ نسبت به برآورد $\kappa_۲$ با سرعت کمتری به مقدار واقعی نزدیک شده است. لذا، انتظار می‌رود با افزایش حجم نمونه برآورد پارامتر اولی با مقدار MSE کمتری مقدار واقعی را برآورد کند. همچنین، به علت اینکه برآورد نسبت آمیختگی (p) با MSE بسیار پایینی برآورد شده است، انتظار می‌رود که الگوریتم EM با خطای ناچیزی بتواند داده دورافتاده را تشخیص دهد.

جدول (۱): برآورد پارامترهای مدل آمیزه فون‌میزس بر اساس الگوریتم EM، مقادیر داخل پرانتز مقدار MSE را نشان می‌دهد.

پارامتر	مقدار واقعی	n = ۱۰۰	n = ۵۰۰	n = ۱۰۰۰
$\mu_۱$	۱/۵۷	۱/۶(۰/۰۴)	۱/۵۷(۰/۰۰۶)	۱/۵۷(۰/۰۰۲۴)
$k_۱$	۱۰/۳	۱۸/۵۶(۱۲۷/۵۲)	۱۱/۸۶(۱۹/۷)	۱۰/۹۶(۷/۲۵)
$\mu_۲$	۳/۱۴	۳/۱۳(۰/۰۰۱)	۳/۱۴(۰/۰۰۰۱۷)	۳/۱۴(۰/۰۰۰۰۹)
$k_۲$	۱۲/۸	۱۳/۱۲(۵/۲۱)	۱۲/۵۶(۰/۷۱)	۱۲/۶۴(۰/۳۶)
$p_۱$	۰/۰۵	۰/۰۶(۰/۰۰۸)	۰/۰۵(۰/۰۰۰۱)	۰/۰۵(۰/۰۰۰۰۴)

برای اطمینان از تفکیک درست پارامترها، می‌توان از ماتریس درهم‌ریختگی استفاده کرد. با استفاده از ماتریس درهم‌ریختگی معیارهایی برای اعتبار سنجی تشخیص خوشه شامل داده‌های

دورافتاده برای حجم نمونه‌های متفاوت در هر تکرار شبیه سازی، محا سبه شده ا ست. میانگین این معیارها به ازای ۵۰۰ بار تکرار در جدول ۲ آورده شده ا ست. در این جدول، معیار در ستی[۱] (دقت خو شه‌بندی داده‌ها)، حسا سیت[۲] (نسبت تشخیص در ست خو شه آلوده)، شیوع[۳] (نسبت وقوع خو شه آلوده) و نرخ مثبت آلوده[۴] (نسبت تشخیص غلط خو شه شامل داده دورافتاده) را با اســتفاده از روش الگوریتم EM نشــان می‌دهند. با توجه به نتایج جدول ۲ ملاحظه می‌شــود که معیار نرخ مثبت آلوده مقدار ۰/۰۰۱ را نشان می‌دهد، به این معنی که الگوریتم EM در ۰/۰۰۱ موارد داده‌های دورافتاده را تشخیص نداده است که این مقدار خطا بسیار ناچیز است.

جدول (۲): میانگین معیارهای اعتبارسنجی تشخیص خوشه مربوط به داده‌های دور افتاده

معیارهای اعتبارسنجی	$n = ۱۰۰$	$n = ۵۰۰$	$n = ۱۰۰۰$
درستی	۰/۹۹	۰/۹۹	۰/۹۹
حساسیت	۰/۹۶	۰/۹۷	۰/۹۸
شیوع	۰/۰۴۷	۰/۰۴۹	۰/۰۵
نرخ مثبت کاذب	۰/۰۰۲	۰/۰۰۱	۰/۰۰۱

معیار درستی مقدار دقت خوشه‌بندی داده‌ها را ۰/۹۹ محاسبه کرده است. این عدد نشان می‌دهد که روش پیشنهادی در این مقاله توانسته است داده‌های دورافتاده را تا حد زیادی به‌درستی تشخیص دهد.

۴-۲-تحلیل یک مثال واقعی

در این بخش داده‌های مربوط به جهت باد و جهت موج که با استفاده از بویه بوشهر و بر اساس درجه طی یک سال اندازه‌گیری شده مورد بررسی قرار می‌گیرد. در واقع این داده‌ها ماهیتاً داده دایره‌ای هســتند. در واقع، هدف بررســی نوع ارتباط بین متغیرهای این داده‌ها اســت. داده‌های مثال مورد مطالعه، مربوط به جهت موج و باد اســت که توســط بویه بوشــهر این داده‌ها در ۲۴ ساعت هر روز و در طول سال ۲۰۰۸ ثبت شدند. از نقطه نظر جغرافیایی، بویه بو شهر در طول جغرافیایی ۵۰/۴۹ و عرض جغرافیایی ۲۸/۵ قرار دارد.

۱ - Accuracy

۲ - Sensitivity

۳- Prevalence

٤- False positive rate

داده‌های دایره‌ای این مثال همگی برحسب درجه هستند. اما، برای سهولت محاسبات و هم‌چنین جلوگیری از نمادگذاری درجه، تمامی داده‌ها به رادیان تبدیل شدند. علاوه بر این، برای از بین بردن اثر روز تمامی داده‌های ثبت شده در ساعت ۱۶:۳۰ در نظر گرفته شد. یکی از مفروضات اولیه رگرسیون دایره‌ای استقلال آماری بین مشاهدات است. از آنجائی که انتظار می رود داده‌های روزهای متوالی به طریقی همبستگی داشته باشند سعی شد همبستگی حداقل یک روزه داده‌ها بررسی شود. پی مقدار آزمون همبستگی پیرسون برای تأخیر یک برابر $۱/۴۸\times۱۰^{-۱۲}$، تأخیر دو برابر $۲/۳\times۱۰^{-۳}$، تأخیر سه برابر $۱/۲\times۱۰^{-۳}$ و تأخیر چهار برابر ۰/۲۲ به دست آمد. این مقادیر نشان می‌دهد که از تأخیر سه به بعد بین داده‌ها همبستگی وجود ندارد. به عبارتی دیگر، همبستگی داده‌ها بعد از گام چهارم از بین می‌رود. این نتیجه نیز در شکل ۱ مشاهده می‌شود. از آنجائی که تعداد کل داده‌ها در ساعت موردنظر ۳۶۰ عدد بود، با انتخاب هر چهارمین مشاهده تعداد داده‌ها به ۹۰ کاهش یافت. بعد از تحلیل مقدماتی برای پالایش داده‌ها یکی از علاقه‌مندی‌های محاسبه خلاصه‌های آمار دایره‌ای برای این داده‌ها است.

میانگین، میانه، دامنه و واریانس دایره‌ای به‌اضافه میانگین طول برآیند برای داده‌های جهت موج و جهت باد محاسبه و نتایج در جدول ۳ آورده شده است.

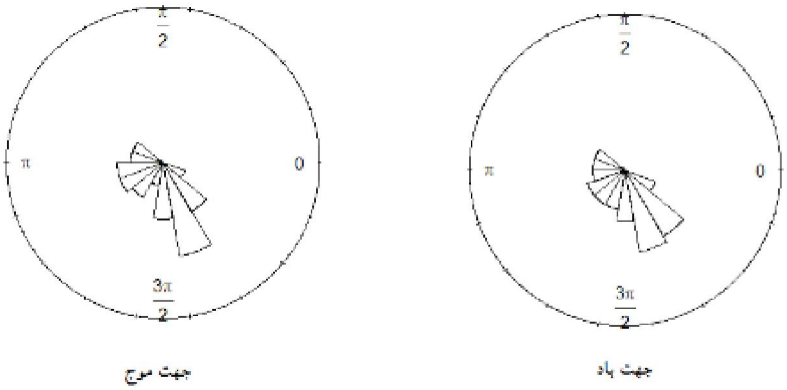
جدول (۳): تحلیل توصیفی داده‌های جهت باد و جهت موج (برحسب رادیان)

متغیر	میانگین	میانه	دامنه	واریانس	میانگین طول برآیند
جهت موج	-۱/۶۷	-۱/۳۷	۲/۹	۰/۳	۰/۷
جهت باد	-۱/۵	-۱/۲۶	۳/۱۷	۰/۲۹	۰/۷

در ستون اول و دوم معیارهای مرکزی این دو متغیر محاسبه شده است. با توجه به مقدار میانگین دایره‌ای، جهت باد و موج غالب منطقه ساحلی بوشهر در فصل به سمت جنوب بوده است. در ستون‌های بعدی جدول ۲ معیارهای پراکندگی آورده شده است. کوچک‌ترین کمانی که شامل همه مشاهدات نمونه‌ای شود، دامنه دایره‌ای است که مقادیر آن برای هر دو متغیر جهت موج و باد در دو فصل در ستون سوم آمده است. با توجه به اینکه واریانس دایره‌ای جهت موج فصل پاییز به یک نزدیک‌تر است پراکندگی بیشتری نسبت به جهت‌های دیگر را دارد.

نمودار گل‌سرخی[۱] یا در اصطلاح جغرافیا گلباد برای هر دو متغیر پاسخ و متغیر تبیینی در شکل ۱ نشان داده شده است. به‌وضوح می‌توان جهت غالب و پراکندگی برای متغیرهای فصل را در این شکل مشاهده کرد.

۱-Rose diagram



شکل (۱): نمودار گلباد جهت موج و باد فصل

برای این امر، با پیروی از [۷]، ابتدا آماره $AICC$ دو مدل رگرسیون دایره‌ای [۵] و [۹] باهم مقایسه می‌شوند. مدلی که مقدار

$$AICC = ۲n \log I_۰(\hat{\kappa}) - ۲n\hat{\kappa} + \frac{n(n+l)}{n-l-۲}$$

کمتری داشته باشد، برازش بهتری به داده‌ها ارائه می‌کند. در این آماره n حجم نمونه و l تعداد پارامترهای برآورد شده هستند. برای مدل [۵] این آماره، ۱۱۸/۸۵- و برای مدل [۹]، مقدارش ۱۱۳/۴۸- به دست آمدند. در نتیجه، باید مدل رگرسیون دایره‌ای JS به داده‌ها برازش داده شود. در برازش این مدل رگرسیونی دایره‌ای، جهت باد به‌عنوان متغیر تبیینی و جهت موج به‌عنوان متغیر پاسخ در نظر گرفته می‌شوند. پارامترهای مدل با فرض اینکه درجه چند جمله‌ای مثلثاتی برابر یک است $(m = ۱)$ به‌صورت زیر برآورد شدند:

$$\hat{A}_۰ = -۰/۱۶۳,\ \hat{A}_۱ = ۰/۸۵۲,\quad \hat{B}_۱ = -۰/۰۷۱,$$
$$\hat{C}_۰ = -۰/۲۷۰,\ \hat{C}_۱ = -۰/۲۸۹,\ \hat{D}_۱ = ۰/۵۸۱.$$

بعد از برآورد پارامترها، توابع $g_۱(x)$ و $g_۲(x)$ به‌صورت

$$\hat{g}_۱(x) = -۰/۱۶۳ + ۰/۸۵۲\cos x - ۰/۰۷۱\sin x$$
$$\hat{g}_۲(x) = -۰/۲۷۰ - ۰/۲۸۹\cos x - ۰/۵۸۱\sin x$$

برآورد شدند. با توجه به اینکه p-مقدار آزمون $\chi^۲$ (برای نحوه محاسبه این آزمون به [۵] مراجعه شود) با سه درجه آزادی برای $\hat{g}_۱(x)$ و $\hat{g}_۲(x)$ به ترتیب ۰/۲۳ و ۰/۹۲ است، نیازی

به افزایش در جه مثل شاتی از یک به‌دو نیســـت. بعد از برآورد $g(x)$ ها، مدل [۵]، یعنی $E(e^{iy}\mid x)=\rho(x)e^{i\mu(x)}=g_1(x)+ig_2(x)$ محاسبه شد، که برای این منظور داریم:

$$\hat{\mu}(x)=\begin{cases}\tan^{-1}\left(\dfrac{-0/270-0/289\cos x_j+0/581\sin x_j}{-0/163+0/852\cos x_j-0/071\sin x_j}\right), & 0/89<x_j<3/17\\ \pi+\tan^{-1}\left(\dfrac{-0/270-0/289\cos x_j+0/581\sin x_j}{-0/163+0/852\cos x_j-0/071\sin x_j}\right), & x_j\le 0/89 \ \ or \ \ 3/17\le x_j\le 2\pi\end{cases}$$

و

$$\hat{\rho}(x)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}^2(x_i)}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}g_1^2(x_i)+g_2^2(x_i))}=0/94.$$

اکنون از رویکرد مدل آمیزه‌ای از توزیع‌های فون‌میزس برای تعیین داده‌های دورافتاده اســتفاده می‌کنیم. با فرض اینکه مانده‌های مدل رگرســیونی JS دارای دو خوشــه هســتند، با اســتفاده از الگوریتم EM به تشــخیص داده‌های دور افتاده پرداختیم. با اســتفاده از این الگوریتم، ملاحظه شد که تنها مشاهده شماره ۲۵ در یک خوشه قرار گرفت و لذا می‌توان آن را به‌عنوان داده دور افتاده در نظر گرفت. این مشاهده مربوط به‌روز جمعه ۲۰ اردیبهشت سال ۱۳۸۷ است.

۵-بحث و نتیجه‌گیری

برخورد با داده‌های دور افتاده در یک تحلیل آماری از مســائل بحث‌برانگیز در آمار اســت. این موضــوع در داده‌های دایره‌ای از اهمیت دوچندان برخوردار اســت، چرا که ماهیت تناوبی بودن داده‌ها بر پیچیدگی موضــوع می‌افزاید. در این مقاله رویکرد جدیدی برای برخورد با این مطلب پیشــنهاد شــد. یکی از مدل‌های رگرســیونی دایره‌ای مطالعه و اســتفاده از آن مدل برای تعیین نقاط دورافتاده دایره‌ای تشــریح شــد. برای ارزیابی عملکرد روش‌های پیشــنهادی مطالعه شبیه‌سازی انجام گرفت. با توجه به نتایج شبیه‌سازی ملاحظه شد که تشخیص داده دورافتاده بر اساس مدل آمیزه‌ای از توزیع فون میزس با روش الگوریتم EM خطای کمی به همراه دارد.

تعمیم تعداد خوشه‌ها از دو به مقادیر بیشتر می‌تواند موضوع مناسبی برای تحقیقات آتی در این حوزه باشــد. به‌علاوه، انتظار می‌رود اســتفاده از روش بیزی در مدل آمیزه فوق‌میزس، میزان خطای برآورد پارامترهای تمرکز خو شه‌ها را به‌اندازه کافی کاهش دهد. اما برر سی این مو ضوع نیاز به تحقیقات وسیع‌تری دارد. در بررسی مثال واقعی سعی شد همبستگی زمانی بین داده‌ها به طریقی از بین برود، تا بدین طریق فرض ا ستقلال م شاهدات م صداق دا شته با شد. پی شنهاد می‌شــود برای مطالعات آینده و بررســی دقیق‌تر این داده‌ها، از تحلیل ســری‌های زمانی دایره‌ای

استفاده شود. بحث راجع به تشخیص داده‌های دور افتاده در سری زمانی دایره‌ای نیز می‌تواند موضوع جالبی برای تحقیقات آتی باشد.

منابع

[1] Mardia, K. V. (1972), *Statistics of Directional Data*, London: Academic Press.

[2] Jammalamadaka, S. R. and SenGupta, A. (2001), *Topics in Circular Statistics*, Singapore: World Scientific.

[3] Gould, A. L. (1969), A Regression Technique for Angular Variates, *Biometrics*, **25**, 683-700.

[4] Laycock, P. J. (1975), Optimal Design: Regression Models for Directions, *Biometrika*, **62**, 305-311.

[5] Sarma, Y., and Jammalamadaka, S. (1993), Circular Regression, *Statistical Science and Data Analysis*, 109-128.

[6] Rivest, L. P. (1997), A Decentred Predictor for Circular-Circular Regression, *Biometrika*, **84**, 717-726.

[7] Lund, U. (1999), Least Circular Distance Regression for Directional Data, *Journal of Applied Statistics*, **26**, 723-733.

[8] Downs, T. D. and Mardia, K. V. (2002), Circular Regression, *Biometrika*, **89**, 683-697.

[9] Hussain, A. G., Fieller, N. R. J. and Stillman, E. C. (2004), Linear Regression for Circular Variables with Application to Directional Data, *Journal of Applied Science and Technology*, **9**, 1-6.

[10] Hussain, A. G., Abdullah, N. A. and Mohamed, I. (2010), A Complex Linear Regression Model, *Sains Malaysian,* **39**, 491–494.

[11] Collett, D. (1980), Outliers in Circular Data, *Journal of Applied Statistics*, **29**, 50-57.

[12] Bagchi, P. and Guttman, I. (1990), Spuriosity and Outliers in Directional Data, *Journal of Applied Statistics*, **17**, 341-350.

[13] Abuzaid, A. H., Mohamed, I. B. and Hussain, A. G. (2009), A New Test of Discordancy in Circular Data, *Communications in Statistics-Simulation and Computation*, **38**, 682-691.

[14] Abuzaid, A., Mohamed, I., Rambli, A. and Hussain, A. G. (2011), COVRATIO Statistic for Simple Circular Regression Model, *Chiang Mai Journal of Science*, 38, 321-330.

[15] Ibrahim, S., Rambli, A., Hussain, A. G. and Mohamed, I. (2013), Outlier Detection in a Circular Regression Model Using COVRATIO Statistic, *Communication in Statistics-Simulation and Computation*, 42, 2272-2280.

[16] Abuzaid, A. H., Hussain, A. G. and Mohamed I. B. (2013), Detection of Outliers in Simple Circular Regression Models Using the Mean Circular Error Statistic, *Journal of Statistical Computation and Simulation*, **83**, 269-277.

[17] Mardia, K. V. and Jupp, P. (2000), *Directional Statistics*, New York: John Wiley.

[18] McLachlan G. J. and Krishnan T. (2008) *The EM Algorithm and Extensions*, New York: John Wiley.

Circular Outliers Detection Using a Mixture of Von Mises Distributions

Khadigeh Abdi, Moussa Golalizadeh, and Taban Baghfalaki

Department of Statistics, Tarbiat Modares University, Tehran, Iran.

Abstract

Circular data are typical examples of directional data with specific systematic period. Because the presence of outliers leads to invalid statistical inferences on the parameters of the circular regression models, their prevalence in analyzing these models requires particular attentions. There are different approaches for modeling the structure of the data set with outliers in which using the mixture models is among the most important ones. As a new idea, to study the methods in determining the outlier data and to employ the mixture model of Von Mises distribution are considered in this paper. The EM algorithm is utilized for estimating the parameters of these models. The performance of the proposed models is investigated using some simulation studies and then the confusion matrix is considered to evaluate the accuracy of fitting models . Also, the proposed methods are applied to analysis a real data set, related to the wave directional analysis in the Bushehr province in Iran.

: ***Keywords***Directional data; Outlier points; Von Mises distribution; Mixture models; *EM* algorithm.

Mathematics Subject Classification (2010): 62H11, 62J20.

کد الگوریتم *EM* برای برآورد پارامترهای مدل

```
iteration=500
 M=matrix(۱,nrow=iteration,ncol=۵)
colnames(M)=c("mu۱","mu۲","kappa۱","kappa۲","p1")
  for(kk in ۱:iteration) {
n=۱۰۰۰۰
h= ۱۰۰۰۰
k=2
meanobs۱<-۰,۵*p1
meanobs۲<-p1
kappaobs۱<-۱۰,۳
kappaobs۲<-۱۲,۸
pobs<-۰,۰۱
z<-rbinom(n,۱,pobs)
randata۱<-rvm(n,meanobs۱,kappaobs۱)
randata۲<-rvm(n,meanobs۲,kappaobs۲)
theta<-(z*randata۱+(۱-z)*randata۲)%%(۲*pi)
 mu=kappa=rep(۰،۲)
Results=matrix(۱,h,۵)
colnames(Results)=c("mu۱","mu۲","kappa۱","kappa۲","p1")
Results[۱,]=c(۰,۰۳۵*pi,۱,۰۱*pi,۱۰,۳،۱۰,. ۱)
      for(c in ۲:h)
    }
p=c(Results[(c-۱),۵],۱-Results[(c-۱),5)]
mu=Results[(c-۱),1:2]
  kappa=Results[(c-۱),3:4]
f=matrix(c(rep(۰,n*k)),n,k)
   for(j in ۱:k){
    f[,j]<-dvm(theta,mu[j],kappa[j])
           }
tt<-matrix(c(rep(۰,n*k)),n,k)
   for(i in ۱:n)
      {        for(j in ۱:k)
         {    tt[i,j]<-(p[j]*f[i,j])/(p[۱]*f[i,۱]+p[۲]*f[1,۲)]
   }   }
   for(j in ۱:k)
    } sorat<-sum(tt[,j]*sin(theta))
     makhrj<-sum(tt[,j]*cos(theta))
```

```
     mu[j]<-atan(sorat/makhrj)
if(makhrj<۰)
{
mu[j]=pi+atan(sorat/makhrj)
}else{
mu[j]=ifelse(sorat>۰,atan(sorat/makhrj),۲*pi+atan(sorat/makhrj))
}
            }
     Results[c,۱:۲]=c(mu[۱],mu[2])
d=c()
   for(j in ۱:k)
    {                d=(sum(tt[,j]*cos(theta-mu[j])))/sum(tt[,j])
       kappa[j]<-(۰,۵۳-۱,۲۸*d^۲)*tan((pi*d)/2)
   {                Results[c,۳:۴]=c(kappa[۱],kappa[2])
    Results[c,۵]<-sum(tt[,۱])/n
    p=c(Results[c,۵],۱-Results[c,5])
if (t(Results[c,]-Results[(c-۱),])%*%(Results[c,]-Results[(c-
۱),])<۰,۰۰۰۰۰۱)(break)
print(c )}
M[kk,]=Results[c,]
 print(kk) }
M[kk,]
 a<-c()
zz<-max.col(tt)
  for(i in ۱:length(zz))
       {      if (zz[i]==۱) a=c(a,zz[i])
          i=i+۱      }
     index=۱:n
     index[zz==۱]
       index[zz==2]
m=c(meanobs۱, meanobs۲, kappaobs۱, kappaobs۲, pobs)
MSE<-matrix(۱,nrow=۱,ncol=۵)
colnames(MSE)=c("MSE.mu۱","MSE.mu۲","MSE.kappa۱","MSE.ka
ppa۲",
"MSE.p1")
     for (j in ۱:iteration)
         {    for (i in ۱:۵)
        {       MSE[i]=(۱/j)*(sum((M[,i]-m[i])^۲))
              }
}
        MSE
```